In einem einjährigen Pilotprojekt digitalisierte die Badische Landesbibliothek (BLB) die in der BLB erhaltenen Inkunabeln aus der ehemaligen Bibliothek des Klosters Reichenau und erschloss diese in maschinenlesbarer Form. Mit einem Korpus von über 70.000 Images setzte das Projekt neue Maßstäbe in der Volltexterschließung einer Materialgattung, die für die computergestützte Texterkennung lange als ungeeignet galt. Während im eigentlichen Texterkennungsvorgang sehr gute Ergebnisse erzielt werden konnten, erwies sich die vorangehende Segmentierung der Seitendigitalisate in Textregionen und Zeilen als fehleranfällig: Hier mussten der angestrebte Automatisierungsgrad und die Qualität der Erschließung gegeneinander abgewogen werden. Die Badische Landesbibliothek stellt die im Rahmen des Projektes erarbeiteten Trainingsdaten und Texterkennungsmodelle unter einer Creative-Commons-Lizenz zur Verfügung.

In a one-year pilot project, the Badische Landesbibliothek (BLB) digitised the incunabula from the former library of the Reichenau monastery preserved at the BLB and rendered them accessible in machine-readable form. Based on a corpus of over 70,000 images, the project set new standards in the full-text processing of a type of material that had long been considered unsuitable for computer-aided text recognition. While very good results were achieved in the actual text recognition process, the preceding segmentation of the digitised pages into text regions and lines proved prone to error. Here, the desired degree of automation had to be weighed against the quality of the resulting texts. The Badische Landesbibliothek makes the training data and text recognition models developed in the project available under a Creative Commons licence.

KATHARINA OST

Möglichkeit und Grenzen einer groß angelegten Volltexterschließung von Inkunabeln mit Transkribus

Ziele und Umfang des Projekts

Anlässlich des Jubiläums zum 1.300-jährigen Bestehen des Benediktinerklosters auf der Reichenau (Gründung 724 n. Chr.) digitalisierte die Badische Landesbibliothek mit Förderung durch die Stiftung Kulturgut Baden-Württemberg die in der BLB erhaltenen Inkunabeln aus der ehemaligen Klosterbibliothek (242 Titel) und erschloss diese in einem ehrgeizigen Pilotprojekt in maschinenlesbarer Form.

Die Badische Landesbibliothek beherbergt neben zahlreichen mittelalterlichen Handschriften auch eine umfangreiche Inkunabelsammlung Reichenauer Provenienz. Diese Drucke gelangten teils durch Erwerbungen für den Privat- und Gemeinschaftsbesitz der Konventualen, teils durch Überlassungen aus den Bibliotheken dem Kloster verbundener Personen in die Reichenauer Klosterbibliothek[1] und wurden 1804 im Zuge der Säkularisation in die damalige großherzoglich-badische Hofbibliothek in Karlsruhe verbracht. 1989 wurde dieser Inkunabelbestand von Felix Heinzer in einer ersten Studie erschlossen.[2] Seitdem konnten aber über 40 weitere Titel Reichenauer Provenienz im Bestand der Badischen Landesbibliothek identifiziert werden.

So ergab sich ein Projektkorpus von 242 Einzeldrucken (219 Inkunabeln, 23 Frühdrucke) in 130 Bänden.[3] Die große Mehrheit der Titel (226) ist in lateinischer Sprache verfasst, hinzu kommen elf deutsche Drucke und fünf lateinisch-deutsche Bilinguen (z. B. Glossare). Das Korpus trägt ein deutliches regionales Gepräge und ein klares thematisches Profil: Die meisten Drucke entstammen dem Gebiet am Oberrhein und dem schwäbischen Raum, der am häufigsten vertretene Druckort ist allerdings Venedig (vor Straßburg und Basel).[4] Dies ist einem inhaltlichen Schwerpunkt in den Bereichen Theologie und Recht geschuldet. Wie aber zu sehen sein wird (s. Abschnitt »Ergebnisse der Texterkennung«), kann das Korpus als Anwendungsfall automatischer Texterkennung nichtsdestotrotz eine gewisse Repräsentativität für lateinische Inkunabeln beanspruchen.

Das einjährige[5] Pilotprojekt zielte nicht nur darauf, eine Sammlung von hoher landesgeschichtlicher Bedeutung besser zu erschließen, sondern sollte auch erproben, inwieweit eine Klasse historischer Dokumente, die für die automatische Texterkennung lange als ungeeignet galt, durch moderne Methoden maschinellen Lernens unter beschränktem Mitteleinsatz bearbeitbar wird. Mit einem Korpus von über 70.000 Images setzte das

https://doi.org/10.3196/1864295024713338 - am 13.01.2026, 14:16:01. https://www.inlibra.com/de/agb - Open Access -

Reichenau-Projekt quantitativ neue Maßstäbe für die computergestützte Volltexterschließung von Inkunabeln und sammelte zahlreiche praktische Erkenntnisse in der Bearbeitung einer Materialgattung, die von dem Verbundprojekt OCR-D nicht abgedeckt wird.[6]

Vorüberlegungen

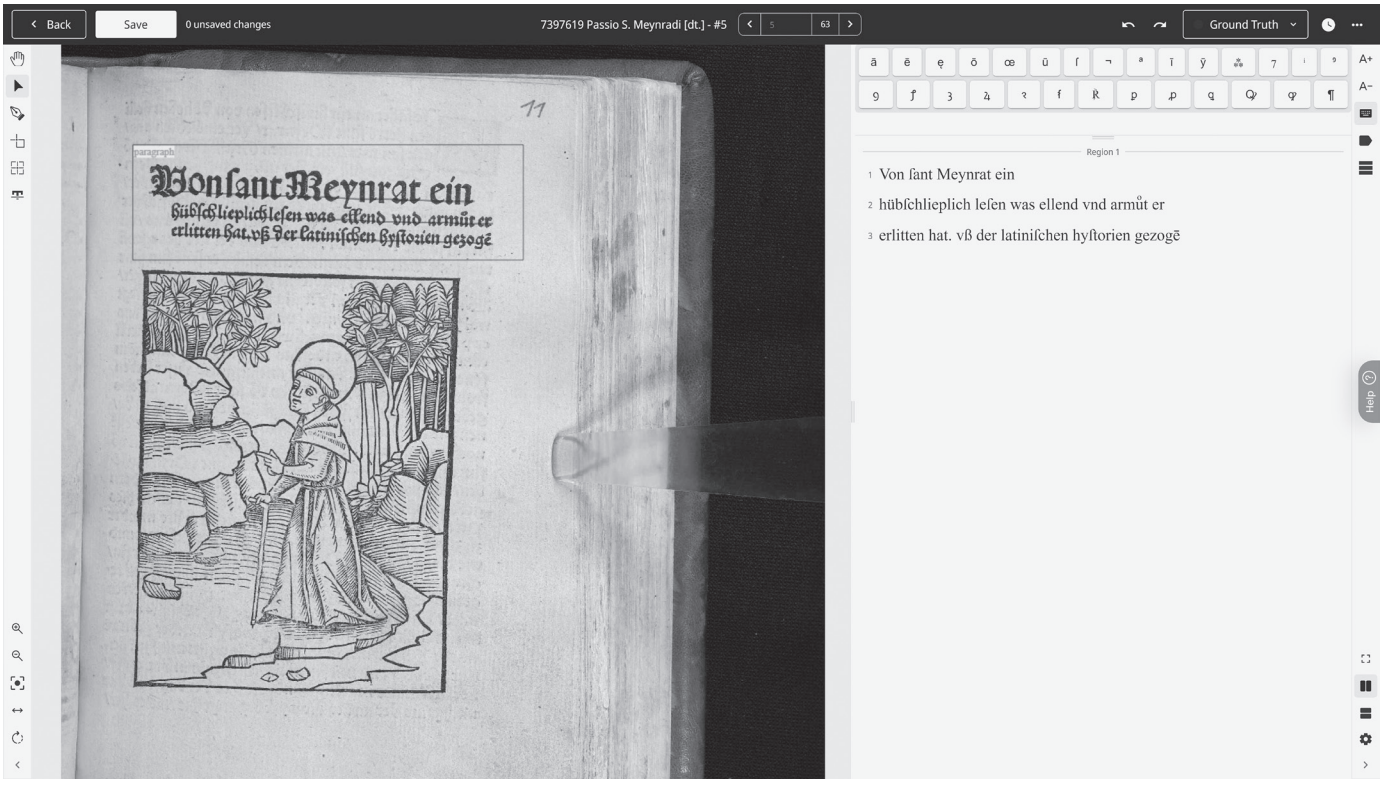
Wahl eines Texterkennungssystems

Im Gegensatz zu modernen Drucken stellten Inkunabeln für herkömmliche Texterkennungssysteme ein großes Problem dar:[7] Diese segmentierten die Bilddaten in einzelne Zeichen, die sie anschließend zu identifizieren versuchten – ein Ansatz, der angesichts der Vielfalt druckereispezifischer Schrifttypen, der Verwendung eines umfangreichen Abbreviaturensystems, der Ergänzung handschriftlicher Initialen und Rubriken sowie der sehr unterschiedlich ausfallenden Druckqualitäten und Erhaltungszustände keine zufriedenstellenden Ergebnisse erzielen konnte.

Erst die Verwendung künstlicher neuronaler Netze (CNN×RNN-Architektur), die jeweils einen kompletten Zeilenkontext berücksichtigen, ermöglichte lernfähige und fehlertolerante Texterkennungssysteme, die an die spezifischen Anforderungen von Inkunabeln und Frühdrucken angepasst werden können. Mittlerweile basieren zwar mehrere Systeme für die automatische Texterkennung auf dieser Technologie,[8] deren Anwendung erfordert aber eine Benutzeroberfläche, die die manuelle Korrektur des Trainingsmaterials unterstützt und Funktionen zur Layoutsegmentierung integriert. Für die Arbeit an historischen Dokumenten bieten sich hier derzeit drei Optionen an: eScriptorium (Université Paris Sciences et Lettres), OCR4all (Universität Würzburg) und Transkribus (READ COOP Kooperative).[9]

eScriptorium und OCR4all werden unter einem Open-Source-Modell entwickelt und ermöglichen ihren Nutzer*innen einen uneingeschränkten Zugriff auf die von ihnen trainierten Texterkennungsmodelle, erfordern allerdings beide den Betrieb eines lokalen Servers. Neben dem anfallenden Administrationsaufwand ist dabei zu berücksichtigen, dass das Training größerer Texterkennungsmodelle nur mit teurer Hardware sinnvoll zu bewerkstelligen ist, die jedoch von einzelnen Institutionen kaum dauerhaft ausgelastet werden kann.[10] Perspektivisch wäre dieses Problem über die Bildung regionaler Verbünde lösbar, die die entsprechenden Hardware-Ressourcen gemeinsam vorhalten und GLAM-Einrichtungen zur Verfügung stellen.

Für das hier vorgestellte Reichenau-Projekt als einem in Dauer und Umfang begrenzten Pilotprojekt fiel die Wahl auf Transkribus. Da Transkribus sämtliche Dokumente auf den Servern der READ-COOP-Kooperative in Innsbruck verarbeitet, fällt zwar eine über abonnierte Seitenkontingente abgebildete Nutzungsgebühr an, es



1 Die Transkribus Web App. Geöffnet ist das Digitalisat der Passio S. Meynradi [dt.], Basel: Michael Furter, um 1495, BLB Dg 198a, fol. a1r

https://doi.org/10.3196/1864295024713338 - am 13.01.2026, 14:16:01. https://www.inlibra.com/de/agb - Open Access -

entsteht aber kein Aufwand im Betrieb eigener Server. READ COOP bietet zwei Benutzeroberflächen an, die auch ohne vertiefte IT-Kenntnisse verwendbar sind: Eine über den Browser zugängliche Web-App (https://app.transkribus.eu/) und eine plattformübergreifend verfügbare Client-Software (»Transkribus Expert Client«, vgl. Abb. 1), deren Weiterentwicklung jedoch eingestellt wurde.[11] Zum aktuellen Zeitpunkt empfiehlt sich die parallele Verwendung beider Zugänge: Zum einen sind wesentliche neue Features (z. B. Field-Modelle) nur über die Web-App verfügbar, zum anderen fehlen dieser jedoch noch zahlreiche Funktionen, die das Arbeiten an größeren Projekten erleichtern (z. B. die Möglichkeit, Transkripte zu importieren). Es steht zu hoffen, dass die Weiterentwicklung der Web-App mittelfristig ein durchgängiges Arbeiten in dieser Umgebung ermöglichen wird; für das Reichenau-Projekt wurde aber überwiegend der Expert Client verwendet.

Der Einfachheit in Bedienung und Einrichtung steht die Bindung an die Transkribus-Plattform gegenüber, die zwar einen Export der erzeugten Volltexte ermöglicht, den Download der Texterkennungsmodelle selbst aber aus kommerziellen Gründen verwehrt: Diese können ausschließlich auf den Transkribus-Servern genutzt und mit anderen Nutzer*innen der Plattform geteilt werden.

Transkriptionsstandards

Die automatische Texterkennung erfordert die Definition von innerhalb des jeweiligen Projekts verbindlich angewandten Transkriptionsstandards: Die formale Einheitlichkeit der für das Training neuer Texterkennungsmodelle verwendeten Daten bedingt die Qualität der resultierenden Texterkennung,[12] wobei neben der bloßen Textrepräsentation auch die Annotation des Seitenlayouts mit zu bedenken ist.

Im Reichenau-Projekt erschienen hinsichtlich der Definition angemessener Transkriptionsstandards drei Überlegungen von besonderer Bedeutung: die für die automatisch erstellten Transkripte angestrebten Nutzungsszenarien, die für das Projekt verfügbaren Ressourcen und die potenzielle Nachnutzbarkeit der erzeugten Transkripte.

Unter dem Gesichtspunkt der Nachnutzbarkeit stellen Trainingsdaten und automatisch erstellte Transkripte Forschungsdaten dar, die den FAIR-Prinzipien genügen sollten. Die Dimensionen Interoperabilität und Wiederverwendbarkeit hängen dabei nicht allein von der Verwendung offener Dateiformate (ALTO oder PAGE XML) ab, sondern auch von den angewandten Transkriptionsrichtlinien. Für historische Dokumente besteht diesbezüglich kein einheitlicher Standard, mögliche Orientierungspunkte können aber die im Rahmen des OCR-D Projektes definierten Erfassungs-Level[13] sowie die Transkriptionsrichtlinien der CATMuS (Consistent Approaches to Transcribing ManuScripts) Initiative[14] bieten. Werden wie für die Volltexterkennung der Reichenauer Inkunabeln projektspezifische Transkriptionsregeln verwendet,[15] sind diese detailliert zu dokumentieren.

Die für das Reichenau-Projekt verfügbaren Ressourcen bedingten die Entscheidung, die Annotation des Seitenlayouts auf ein absolutes Minimum, nämlich die Markierung möglichst großflächiger Textregionen zu beschränken. Die Markierung von Illustrationen und Schmuckelementen oder auch eine detaillierte Abbildung der Textstruktur durch die Annotation von Paragraphen, Überschriften und Seitenzahlen[16] hätte einen hohen manuellen Arbeitsaufwand erfordert, der wenig zu den angestrebten Nutzungsszenarien beiträgt.

Die Nutzungsszenarien der Reichenau-Volltexte sind einerseits durch die Nutzerinnen und Nutzer der Digitalen Sammlungen der Badischen Landesbibliothek bestimmt: Die Option, neben dem Bilddigitalisat auch auf einen Volltext in gewohnter Antiqua-Type zugreifen zu können, ermöglicht es, z. B. Textteile unkompliziert in andere Anwendungen zu kopieren oder über Volltextsuchen auf Inkunabelinhalte zuzugreifen. Andererseits stellen Volltexte auch ein Rohmaterial dar, das durch Methoden der maschinellen Sprachverarbeitung (Natural Language Processing, NLP) weiter erschlossen werden kann. Der Zielkonflikt zwischen Lesbarkeit und Datenqualität wird in der Frage nach dem Ausschreiben von Abbreviaturen konkret: Prinzipiell können Texterkennungsmodelle darauf trainiert werden, einfache Abkürzungen mit einer gewissen Kontextsensitivität auszuschreiben.[17] Ein solches Ausschreiben erhöht zwar die Lesbarkeit der Texte, führt zugleich aber unsystematische Fehler in den Texterkennungsvorgang ein,[18] die nachträglich nur mit großem Aufwand korrigiert werden können. Das Versprechen, so auch paläografisch unerfahrene Leser*innen für historische Materialien zu gewinnen, wirft die Frage auf, wie man diesen die begrenzte Zuverlässigkeit des Verarbeitungsverfahrens vermitteln kann. Für die Reichenauer Inkunabeln fiel daher die Entscheidung, Abbreviaturen nicht automatisch ausschreiben zu lassen. Die hohe Abkürzungsdichte der Texte würde zu stark fehlerbehafteten Ergebnissen führen, während die Korpusgröße von ca. 70.000 Images einen manuellen Korrekturvorgang nicht zulässt. Ein relativ diplomatischer Transkriptionsansatz ergibt hingegen eine qualitativ hochwertige Datenbasis, wobei die Möglichkeit, die Abbreviaturen in einem separaten Verarbeitungsschritt doch noch aufzulösen, gewahrt bleibt.[19] Allgemein gilt, dass eine nachträgliche Normalisierung des Transkripts oft automatisiert möglich ist, der umgekehrte Weg aber eine komplette Neuerschließung des Quellmaterials erfordert.

Für die zahlreichen einzelnen Transkriptionsentscheidungen im Rahmen der Volltexterschließung der Reichenau-Inkunabeln sei auf die zusammen mit den Trainingsdaten zu veröffentlichende Dokumentation verwiesen. Einen nützlichen Leitfaden für grundsätzliche Aspekte, die bei der Definition projektspezifischer

https://doi.org/10.3196/1864295024713338 - am 13.01.2026, 14:16:01. https://www.inlibra.com/de/agb - Open Access -

Transkriptionsrichtlinien zu bedenken sind, bilden die von Estelle Guéville und David Wrisely formulierten sieben Prinzipien für das Training von Texterkennungsmodellen zur Transkription mittelalterlicher Manuskripte.[20] Beispielhaft seien hier die Empfehlung, Stichproben des gesamten Korpus vor der Festlegung von Transkriptionsrichtlinien zu sichten (Prinzip 2), sowie die Notwendigkeit einer vertieften buchwissenschaftlichen Materialkenntnis (Prinzip 6) genannt.[21]

Arbeitsablauf

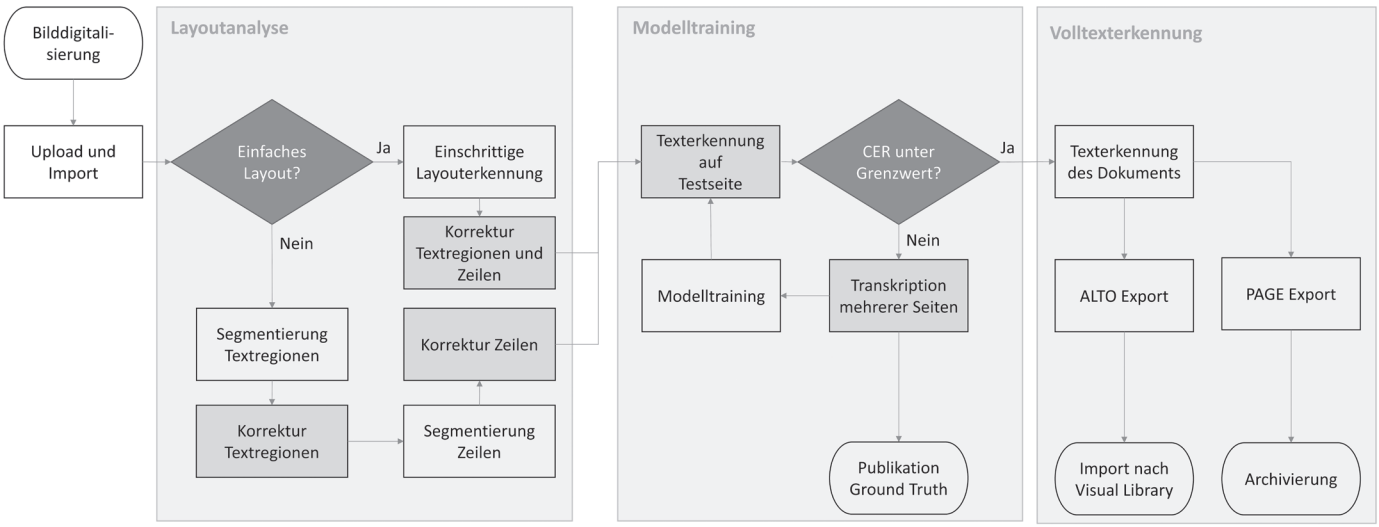
Die Volltexterschließung lässt die bestehenden Arbeitsabläufe der Bilddigitalisierung an sich unberührt und ergänzt diese um einen weiteren Verarbeitungsschritt: Nach der restauratorischen Freigabe der Objekte stellt die Digitalisierungswerkstatt hochauflösende Bilddigitalisate her und speist diese als TIFF-Dateien in das Visual Library System der Badischen Landesbibliothek ein. Es folgen die Strukturannotation und Qualitätskontrolle der Digitalisate, die Anreicherung mit formalen Metadaten im Katalogsystem K10plus und ggf. bereits die Freischaltung für die öffentliche Nutzung. Erst hier setzt die Volltexterkennung an: Indem sie erst nach der Bearbeitung des Digitalisats in Visual Library erfolgt, kann sie auf exportierte Bilddateien zurückgreifen, die in ihren Dateinamen bereits die zugehörige VLID tragen – ein Identifikationsschlüssel, der die automatische Zuordnung der in Transkribus erzeugten Volltexte zu den in Visual Library gespeicherten Digitalisaten ermöglicht. Die so exportierten Bilddateien werden über eine FTP-Schnittstelle auf die Transkribus-Server geladen, um für die Arbeit innerhalb der Softwareumgebung zur Verfügung zu stehen.

Der Texterkennungsprozess selbst erfolgt in drei Schritten (vgl. Abb. 2):

(1) Da die von Transkribus unterstützten Texterkennungsmodelle jeweils den Bildausschnitt einer einzelnen Zeile als Input erwarten, besteht der erste Verarbeitungsschritt darin, die Bilddateien im Zuge einer vorbereitenden Layoutanalyse in Textregionen und Zeilen segmentieren zu lassen. Dieser Vorgang ist relativ fehleranfällig, für die Qualität der anschließenden Texterkennung aber von großer Bedeutung (s. Abschnitt »Ergebnisse der Layouterkennung«). Daher sollten die Ergebnisse der automatisierten Layoutanalyse gesichtet und ggf. manuell korrigiert werden. Diese im Diagramm grau hinterlegten Arbeitsschritte benötigen einen hohen Arbeitsaufwand. Je nach Komplexität des Materials kann es möglich sein, Textregionen und Zeilen in einem einzigen Verarbeitungsschritt segmentieren zu lassen, teils ist es aber unabdingbar, zunächst nur Textregionen erkennen zu lassen, diese zu korrigieren und anschließend die Zeilenerkennung durchzuführen.

(2) Anschließend ist anhand einer exemplarischen Seite aus dem Dokument zu prüfen, ob das bestehende Texterkennungsmodell dessen Schriftbild bereits hinreichend zuverlässig erkennt. Liegt die berechnete Zeichenfehlerrate (Character Error Rate, CER)[22] oberhalb des angestrebten Grenzwerts von 0,6 %, muss eine neue Modellversion trainiert werden: Durch die manuelle Korrektur weiterer Transkripte des fraglichen Dokuments wird zusätzliches Trainingsmaterial erstellt, das in das Training der nächsten Modelliteration miteinfließt. Diese ist wiederum an einer Testseite des Dokuments zu prüfen.

(3) Ist die Zeichenfehlerrate der Stichprobe akzeptabel, wird das Texterkennungsmodell auf das gesamte Dokument angewendet. Eine abschließende manuelle Korrektur der Transkripte erfolgt im Rahmen des hier beschriebenen Projektes nicht mehr. Für den anschließenden Import der Transkripte in die Digitalisierungs-



2 Der Volltexterkennungs-Workflow des Reichenau-Projekts

https://doi.org/10.3196/1864295024713338 - am 13.01.2026, 14:16:01. https://www.inlibra.com/de/agb - Open Access -

plattform Visual Library eignet sich das ALTO-Format: Unter Verwendung dieses Dateiformats erlaubt Transkribus eine automatische Koordinatenberechnung, die es in der Präsentationsoberfläche von Visual Library ermöglicht, Suchergebnisse wortweise auf dem Bilddigitalisat hervorzuheben.[23] Für die Archivierung der Projektergebnisse und die Publikation der Trainingsdaten wird hingegen das von Transkribus auch intern verwendete PAGE XML-Format genutzt.

Ergebnisse

Ergebnisse der Layouterkennung

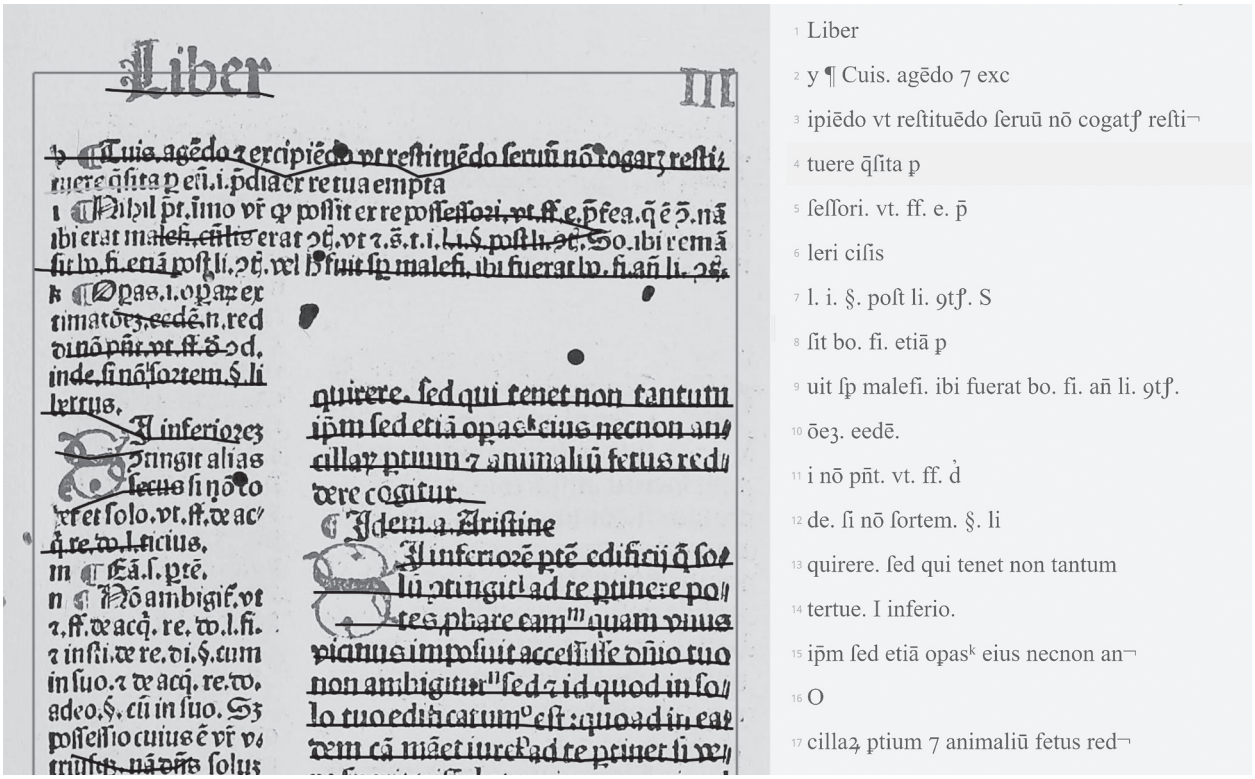
Die automatische Layoutanalyse stellt in der Erfahrung des Reichenau-Projektes die größte Hürde für eine automatische Volltexterschließung von Inkunabeln und Frühdrucken dar. Ihre zentrale Rolle bleibt dadurch verborgen, dass Projekte zur Volltexterkennung in der Regel die Zeichenfehlerrate des Transkripts als wichtigste Qualitätsmetrik nur auf Seiten erheben, deren Layoutrepräsentation (Textregionen und Zeilen) bereits manuell korrigiert wurde. Die so berechneten Ergebnisse sind auf vollständig automatisierte Prozesse aber nur bedingt übertragbar.

Ein Beispiel soll diese Problematik veranschaulichen: Auf einer Stichprobe von fünf Seiten eines typischen juristischen Kommentars (Codex Justinianus mit Glossa ordinaria, Basel: Wenssler, 1487, BLB, Ēb 7) wurden vier unterschiedliche Tiefen der Layoutkorrektur getestet, wobei stets dasselbe Texterkennungsmodell zur Anwendung kam. Das Seitenlayout einer Klammerglosse[24] sowie für Inkunabeln typische Eigenschaften wie handschriftliche Initialen, Rotdruck und Wurmfraß erschwerten die Layout- und Texterkennung.

Layoutanalyse	Character Error Rate	Bag-of-Words F1-Score
Standardeinstellungen	72,6 %	0,498
Field Modell und Universal Lines (Skalierungsfaktor 2)	1,43 %	0,958
Manuelle Korrektur der Textregionen	1,26 %	0,959
Manuelle Korrektur der Textregionen und Baselines	0,74 %	0,968

Tab. 1 Texterkennungsqualität in Abhängigkeit des Grades manueller Layoutkorrektur

Erhoben wurden für das Beispiel jeweils die innerhalb von Transkribus berechnete Zeichenfehlerrate (CER) und ein Bag-of-Words F1-Score,[25] der den Wortbestand von Ground Truth und Transkript ohne Berücksichtigung der Wortreihenfolge evaluiert. Zur Texterkennung wurde das



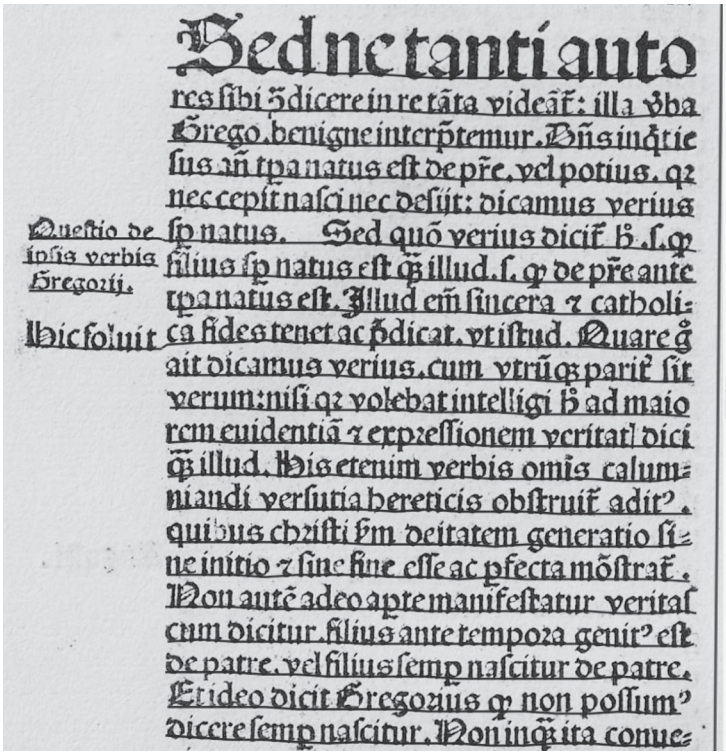
3 Textverluste unter Verwendung der Transkribus-Standardeinstellungen. Horizontale Linien im Digitalisat repräsentieren die im Zuge der automatischen Layoutanalyse erkannten Zeilen. Digitalisat von Codex Justinianus mit Glossa ordinaria, Basel: Wenssler, 1487, BLB, Ēb 7, fol. k9[r]

https://doi.org/10.3196/1864295024713338 - am 13.01.2026, 14:16:01. https://www.inlibra.com/de/agb - Open Access -

im Rahmen des Reichenau-Projektes trainierte Modell für lateinische Inkunabeln in der Version vom 5.2.2024 verwendet, das auf seinem Validierungsset[26] eine CER von 0,58 % und einen BoW F1-Score von 0,977 erreicht.

Die Verwendung der Standardeinstellungen von Transkribus, d. h. die Durchführung einer Texterkennung in der Web-App ohne vorher separate Schritte zur Layoutanalyse vorgenommen zu haben, führte in diesem Beispiel zu großen Textverlusten und einer Zeichenfehlerrate von über 70 %. Um die in einem werkspezifisch konfigurierten, aber automatisch durchgeführten Prozess realistisch erreichbare Texterkennungsqualität abzuschätzen, kam in der zweiten Messung ein sogenanntes Field-Modell für die Textregionen zum Einsatz, das im Rahmen des Reichenau-Projektes spezifisch für Klammerglossen trainiert worden war. Bei der Zeilenerkennung mit dem öffentlichen »Universal Lines«-Modell musste eine automatische Vergrößerung des Eingabebildes zugeschaltet werden, um Textverluste zu vermeiden. Eine manuelle Korrektur der Textregionen erzielte in der dritten Messung nur geringe Verbesserungen, die zusätzliche manuelle Korrektur der Textzeilen (Baselines) konnte aber Textverluste am Zeilenende und Ungenauigkeiten im Bereich der rubrizierten Initialen beheben und somit eine CER von 0,74 % Prozent erreichen. Die in einem automatischen Prozess erreichbare Zeichenfehlerrate von 1,43 % läge in diesem Beispiel also fast doppelt so hoch wie der auf Grundlage manuell korrigierter Layouts berechnete Wert.

Ein typisches materialspezifisches Problem im Bereich der Layoutanalyse stellen gedruckte Marginalien dar. Diese sind in Inkunabeln oft so nah an den Haupttext gedruckt, dass die automatische Layoutanalyse Marginalie und Haupttext auf eine Textzeile setzt, was bei mehrzeiligen Marginalien einen unzusammenhängenden Gesamttext ergibt. Es wäre wünschenswert, die Marginalien der zugehörigen Textstelle möglichst »eng« beizuordnen, sie als Einschübe in den Haupttext einzupflegen (vgl. Abb. 4), dies ist aber mit einem sehr hohen manuellen Arbeitsaufwand verbunden. OCR-D schlägt eine absatzweise Zuordnung der Marginalien vor.[27] Dieses Vorgehen bildet einen attraktiven Kompromiss aus Nähe zum zugehörigen Haupttext und Arbeitsaufwand, der den inneren Zusammenhang beider Textelemente bewahrt – allerdings setzt es typografische Absätze voraus. Diese Annahme ist für Inkunabeln, deren Text oft über Absatzzeichen innerhalb des Blocksatzes strukturiert wird, nicht unbedingt gegeben.[28] Unter Berücksichtigung der begrenzten Projektressourcen fiel für das Reichenau-Korpus die Entscheidung, mehrzeilige Marginalien am Fuß der Seite zu sammeln – die Zuordnung zu der jeweiligen Textstelle geht so zwar verloren, die Lesbarkeit des Haupttextes bleibt aber gewährleistet und der gesamte Wortbestand der Seite wird für die Volltextsuche erschlossen.



44 Sed ne tanti auto
45 res ſibi ꝯdicere in re tāta videāt: illa v̇ba
46 Grego. benigne interp̄temur. Dn̄s inq̇t ie
47 ſus añ tpa natus eſt de pr̄e. vel potius. qꝛ
48 nec cepit naſci nec deſijt: dicamus verius
49 Queſtio de
50 ipſis verbis
51 Gregorij.
52 ſp natus. Sed quō verius dicit ho .ſ. qꝛ
53 filius ſp natus eſt aq3 illud .ſ. qꝛ de pr̄e ante
54 tpa natus eſt. Illud em̄ ſincera 7 catholi¬
55 Hic ſoluit
56 ca fides tenet ac p̄dicat. vt iſtud. Quare og
57 ait dicamus verius. cum vtrūq3 parit ſit
58 verum: niſi qꝛ volebat intelligi h ad maio

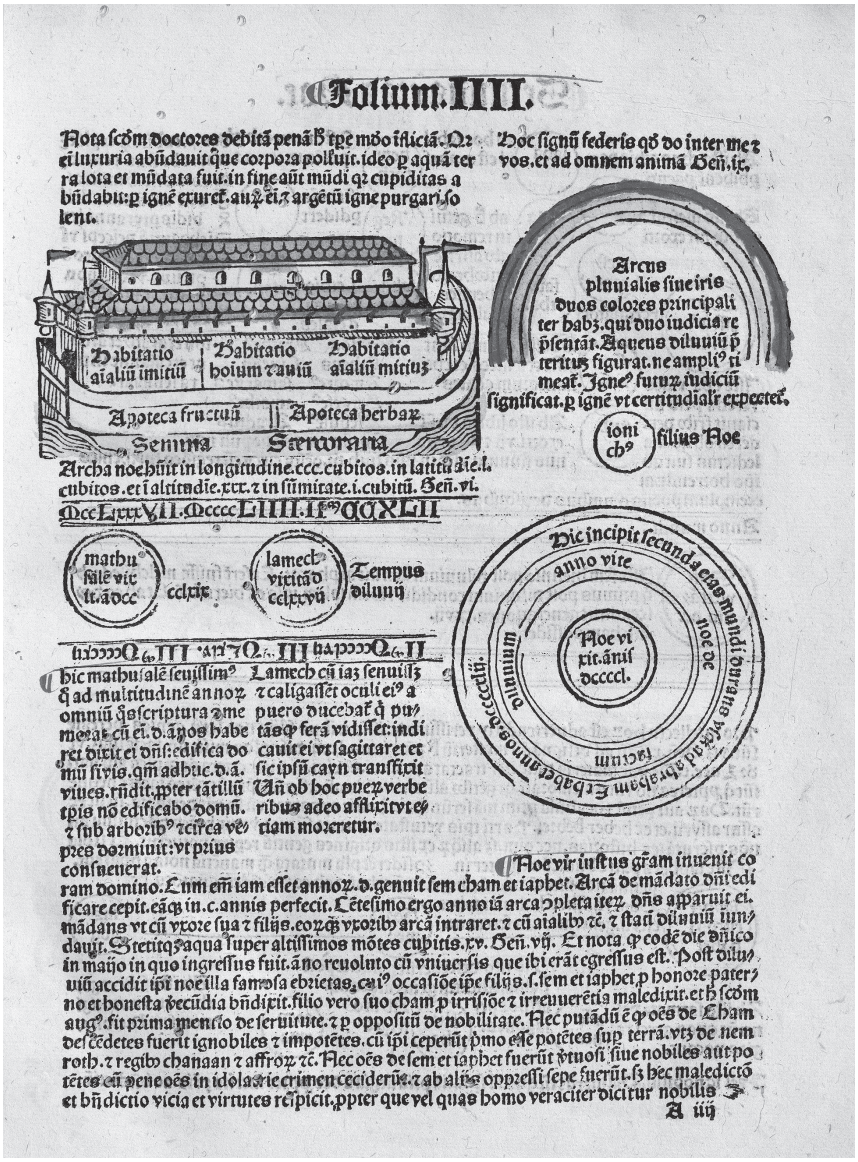
4 Zeilenweises Einpflegen und Tagging von Marginalien am Beispiel eines Abschnitts aus Petrus Lombardus, Sententiarum libri quattuor, Basel: Nikolaus Kessler 1489, BLB, Dg 246, fol. b8v

https://doi.org/10.3196/1864295024713338 - am 13.01.2026, 14:16:01. https://www.inlibra.com/de/agb - Open Access -

Neben Herausforderungen, die ganze Gruppen von Drucken betreffen, sind auch Werke zu berücksichtigen, die aufgrund ihrer aufwendigen Gestaltung nur mit ausgeprägter Materialkenntnis und hohem Arbeitsaufwand für die Texterkennung erschlossen werden können. Beispielhaft sei Werner Rolevincks *Fasciculus temporum* genannt, der sowohl in lateinischen als auch deutschen Ausgaben (1480–1490) im Reichenauer Inkunabelkorpus vertreten ist (vgl. Abb. 5):[29] Diese Weltgeschichte arrangiert Textblöcke, Diagramme und Abbildungen in vertikalen wie horizontalen Ordnungen, um Synchronizität und Diachronie der dargestellten Ereignisse abzubilden, Textzeilen stehen teils gedreht oder fließen kreisförmig.[30] Hier bleibt nur, die Layoutelemente einer jeden der 140 bis 280 Seiten händisch einzuzeichnen. Für das Reichenau-Projekt schien dieser Arbeitsaufwand gerechtfertigt, da gerade aufwendig gestaltete Inkunabeln für die Nutzer*innen der Sammlung von besonderem Interesse sind und dankbare Anknüpfungspunkte beispielsweise für Ausstellungsprojekte bilden. Grundsätzlich stellt sich für massenhafte Digitalisierungs- und Texterkennungsvorhaben jedoch die Frage, ob komplexe »Problemfälle« nicht besser in separate Erschließungsprojekte auszulagern wären.

Ergebnisse des Modelltrainings

Während frühere Projekte zur Texterkennung von Inkunabeln werk- oder druckereispezifische Texterkennungsmodelle trainierten,[31] verfolgt das vorgestellte Projekt einen gemischten Ansatz, in dem der Trainingsdatensatz zwar nach Bedarf um werkspezifisches Material ergänzt wird, die damit trainierten Modelle aber nur nach Sprache differenzieren und nicht auf einzelne Werke oder Schriftbilder zugeschnitten sind.[32]



5 Werner Rolevincks Fasciculus temporum (Straßburg: Johann Prüß, ab 1490, BLB L c̄ 25, fol. 4r) als Beispiel für ein nur manuell erschließbares Seitenlayout

https://doi.org/10.3196/1864295024713338 - am 13.01.2026, 14:16:01. https://www.inlibra.com/de/agb - Open Access -

Dementsprechend wurden drei Texterkennungsmodelle trainiert: Je ein Modell für lateinische, für deutschsprachige und für lateinisch-deutsch-bilinguale Inkunabeln.[33] Eine zu Projektbeginn noch antizipierte Trennung nach Fraktur- und Antiquaschriften erwies sich nicht nur als wenig materialangemessen (z. B. Verwendung von Fraktur als Auszeichnungsschrift in Antiquadrucken), sondern auch als unnötig: Die Mischung beider Schriftfamilien beeinträchtigte die Modellqualität nicht. Eine Trennung nach Sprache wurde hingegen beibehalten, da die Texterkennungsmodelle sprachabhängige Annahmen über die Wahrscheinlichkeit von Zeichenfolgen bilden und das sprachliche Ungleichgewicht des Korpus (228 lateinische, 15 deutsche, 4 bilinguale Titel) deutliche Verschlechterungen bei der Texterkennung deutschsprachiger Werke erwarten ließe. Für das Training eines Modells für bilinguale Drucke wurde dementsprechend ein Datensatz zusammengestellt, der beide Sprachen zu ungefähr gleichen Teilen repräsentiert.

Die für das Modelltraining benötigte Rechenzeit hängt stark von dem Umfang des verwendeten Trainingsmaterials ab und erstreckte sich im Reichenau-Projekt von elf Stunden (Modell für deutschsprachige Inkunabeln vom 25.1.2024, ca. 57.000 Wörter Trainingsmaterial, 250 Trainingsepochen) bis 45 Stunden (Modell für lateinische Inkunabeln vom 2.2.2024, ca. 627.000 Wörter Trainingsdaten, 92 Trainingsepochen). Das Modelltraining ereignet sich aber vollständig auf den READ COOP-Servern und blockiert andere Arbeitsabläufe nicht.

Hinsichtlich des Trainingserfolgs kommt der Qualität des Trainingsmaterials eine kaum zu überschätzende Bedeutung zu. Die in Transkribus trainierten Texterkennungsmodelle neigen zum sogenannten Overfitting, passen sich also zu genau an die Trainingsdaten an: Anstatt wie gewünscht zu generalisieren, reproduziert das Texterkennungsmodell im Training »gesehene« Einzelfälle oft wortwörtlich. So können sich einzelne Transkriptionsfehler bei der anschließenden Anwendung des Modells auf große Textbestände ausbreiten. Zur gründlichen Korrektur der Trainingsdaten ist es daher sinnvoll, innerhalb der Transkribus-Umgebung eine Schriftart zu verwenden, die eine gute Unterscheidbarkeit der verwendeten (Sonder-)Zeichen gewährleistet. Für Inkunabeln und mittelalterliche Handschriften hat sich die Schrifttype Junicode 2 bewährt.[34]

Ergebnisse der Texterkennung

In der Texterkennung erreichte das Reichenau Projekt Zeichenfehlerraten von 0,58 % auf dem lateinischen Validierungsdatensatz (Modellversion vom 2.2.2024, ca. 627.000 Wörter Trainingsdaten). Das mit deutlich weniger Material trainierte Modell für deutschsprachige Inkunabeln erzielte eine CER von 0,55 % (Modellversion vom 25.1.2024, ca. 57.000 Wörter Trainingsdaten), das Modell für bilinguale Texte eine CER von 0,54 % (Modellversion vom 21.2.2024, ca. 113.000 Wörter Trainingsmaterial).[35] Diese Werte sind im Vergleich sehr zufriedenstellend: So bestimmten etwa die Entwickler der Texterkennungssoftware Calamari eine Zeichenfehlerrate von 2 % als Zielgröße »[t]o master OCR on early printed books«.[36] Die erreichte hohe Texterkennungsqualität rechtfertigt somit den Ansatz, die in Transkribus erzeugten Volltexte auch ohne weitere manuelle Korrektur der Öffentlichkeit zur Verfügung zu stellen.

Die Prüfung des Modells für lateinische Inkunabeln auf einer Stichprobe des übrigen Inkunabelbestands der Badischen Landesbibliothek (je eine Seite aus 50 verschiedenen Drucken) erzielte eine Zeichenfehlerrate von 0,73 % und belegt somit die breite Anwendbarkeit dieses Modells auch über das Reichenauer Korpus hinaus.

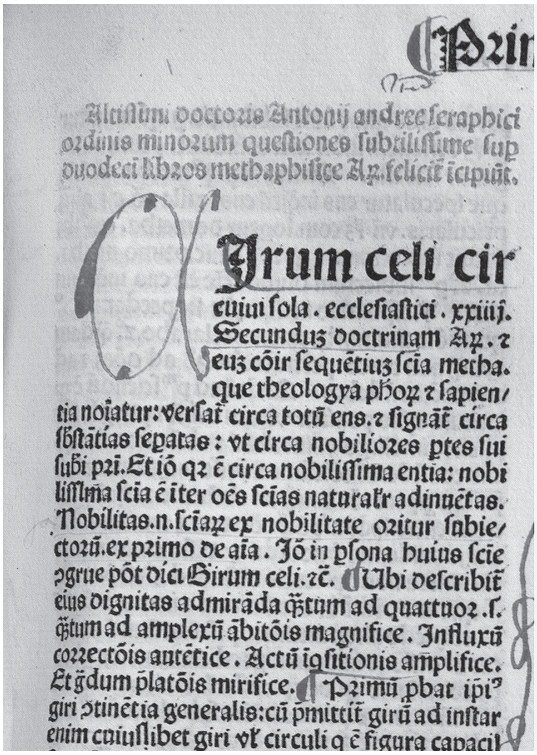
Eine Auswertung der Transkriptionsfehler auf dem lateinischen Validierungsdatensatz des Reichenau-Projektes mit der Dinglehopper-Software des Quratator-Projektes[37] ergab folgende Ergebnisse (152 Seiten, Modellversion vom 2.2.2024):

Die häufigsten Fehler (26,11 %) betreffen die Setzung von Leerzeichen und damit die Wortsegmentierung. Das ist sowohl durch das Material als auch durch den Transkriptionsansatz bedingt: Die Abstände zwischen Buchstaben und Wörtern sind in Inkunabeln oft nicht völlig regelmäßig, sodass sich auch für lesende Menschen ein wenig eindeutiges Bild ergeben kann. Zudem wurden für das 15. Jahrhundert übliche Variationen in der Worttrennung (z. B. *siquidem* / *si quidem*) in den Ground-Truth-Daten nicht normalisiert. Auch die Interpunktion als zweithäufigste Fehlerquelle (11,5 %) betrifft neben der in Inkunabeln notorisch unsystematischen Verwendung von Satzzeichen Fragen der Normalisierung, etwa im Umgang mit Zahlzeichen und Abbreviaturen (».VI.«, ».ſ.«). Die übrigen Fehler sind breit gestreut.

Ground Truth-Daten	**Texterkennung**	Häufigkeit (n=1218)	
Leerzeichen		184	15,11 %
	Leerzeichen	139	11,41 %
·		75	6,16 %
	·	65	5,34 %
ē	e	51	4,19 %
ā	a	49	4,02 %
e	ē	42	3,45 %
ū	u	37	3,04 %
¬		30	2,46 %
	i	30	2,46 %

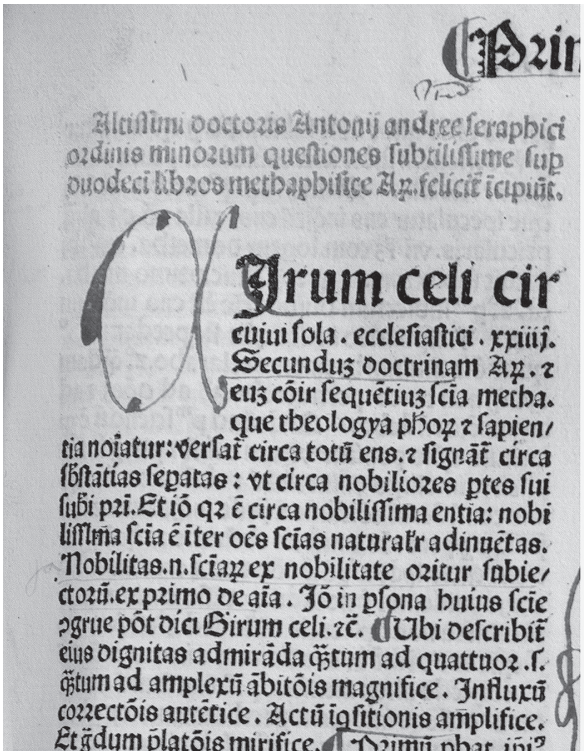
Tab. 2 Die zehn häufigsten Zeichenverwechslungen des Texterkennungsmodells für lateinische Inkunabeln

https://doi.org/10.3196/1864295024713338 - am 13.01.2026, 14:16:01. https://www.inlibra.com/de/agb - Open Access -



1-1 # ¶ Primus.
2-1 # Altiſſimi doctoris Antonij andree ſeraphici
2-2 # ordinis minorum queſtiones ſubtiliſſime ſup
2-3 # duodecī libros methaphiſice A2. felicit īcipiūt.
2-4 # ~~Arum~~ Irum celi cir
2-5 # cuiui ſola. eccleſiaſtici. xxiiij.
2-6 # Secundu3 doctrinam A2. 7
2-7 # ~~Ieu3~~ eu3 cōir ſequētiu3 ſcīa ~~mecha.~~ metha.
2-8 # que ~~theologia~~ theologya pho2 7 ſapien¬
2-9 # tia noīatur: verſat circa totū ens. 7 ſignāt circa
2-10 # ſbſtātias ſepatas: vt circa nobiliores ptes ſui
2-11 # ſubi prī. Et iō q2 ē circa nobiliſſima entia: nobi
2-12 # liſſima ſcīa ē īter oēs ſcīas naturalr adinuētas.
2-13 # Nobilitas .n. ſcīa2 ex nobilitate oritur ſubie¬
2-14 # ctorū. ex primo de aīa. Iō in pſona huius ſcīe
2-15 # 9grue pōt dici Girum celi. 7c. ¶ Ubi deſcribit
2-16 # eius dignitas admirāda qtum ad quattuor .ſ.
2-17 # qtum ad amplexū ābitōis magnifice. Influxū
2-18 # correctōis autētice. Actū īqſitionis amplifice.
2-19 # Et gdum plātōis mirifice. ¶ Primū pbat ipi9
2-20 # giri 9tinētia generalis: cū pmittit girū ad ~~inſta¬~~ inſtar
2-21 # enim cuiuſlibet giri vl circuli q ē figura ~~capaci¬~~ capa
2-22 # ſima ſm geometricos ſcīa metha. oīa ābit. ~~Na¬~~ Nā

6 Ergebnis eines Texterkennungsvorgangs mit typischer Fehlerrate (Modellversion vom 26.10.23, CER 0,62 %, Wortfehlerrate 2,82 %) auf Grundlage des Digitalisats von Antonius Andreas, Quaestiones super XII libros Metaphysicae Aristotelis, Venedig: Antonius de Strata, 1481, BLB, Pb 27a, fol. a2r



Region 1

1 ¶ Primus.

Region 2

1 Altiſſimi doctoris Antonij andree ſeraphici
2 ordinis minorum queſtiones ſubtiliſſime ſup
3 duodecī libros methaphiſice A2. felicit īcipiūt.
4 Arum celi cir
5 cuiui ſola. eccleſiaſtici. xxiiij.
6 Secundu3 doctrinam A2. 7

Region 1

1 timus.

Region 2

1 Altissimi doctoris Antonij andre scraphici
2 ordinis minorum questioneo subtilissime sup¬
3 duodeci libros methaphilice An. felicitur incipiun

7 Texterkennungsmodell des Reichenau-Projekts (rechts oben, Modellversion vom 26.10.2023) im Vergleich mit Noscemus GM 6 (rechts unten)

https://doi.org/10.3196/1864295024713338 - am 13.01.2026, 14:16:01. https://www.inlibra.com/de/agb - Open Access -

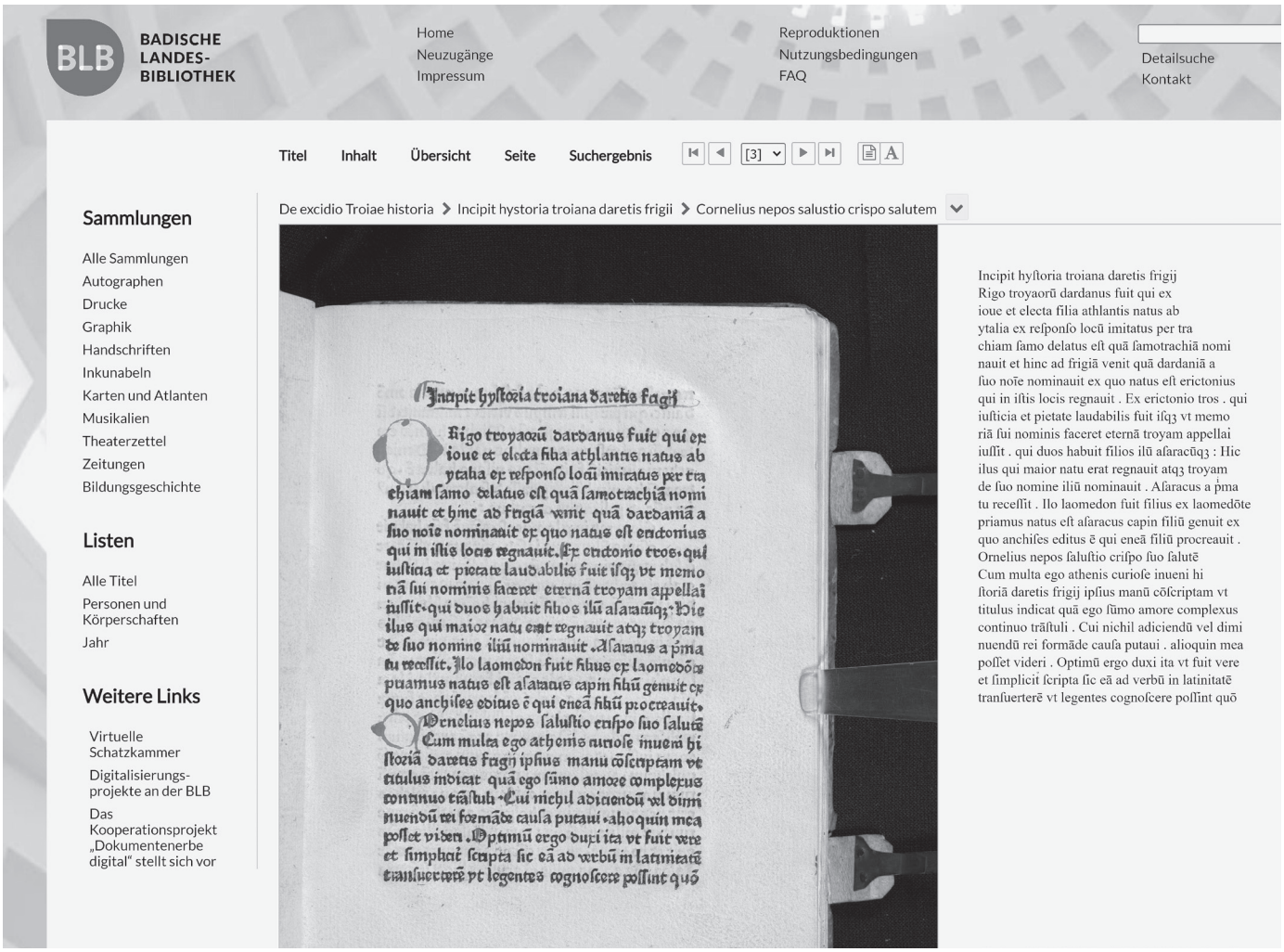
Besonders häufig scheinen Texterkennungsfehler am rechten Seitenrand aufzutreten (vgl. Abb. 6). Dies mag teils Ungenauigkeiten in der Zeilensegmentierung geschuldet sein, teils sind die Fehler aber auch durch die Formenvielfalt der im 15. Jahrhundert verwendeten Trennzeichen erklärbar. Da die sprachverarbeitenden Anteile des Texterkennungsmodells (BLSTM-Ebenen) den linken und rechten Kontext der jeweils betrachteten Stelle mitberücksichtigen, fällt ihre Leistung am Textrand, an dem sie lediglich einen einseitigen Kontext zur Verfügung haben, schwächer aus.

Ein ebenfalls inkunabelspezifisches Problem stellt die Eigenschaft von Transkribus-Modellen dar, die verwendeten Bilddaten in Graustufen oder eine Schwarz-Weiß-Darstellung zu konvertieren. Hierdurch sind die Texterkennungsmodelle grundsätzlich weder in der Lage, handschriftlich eingetragene Rubrizierungen vom (schwarz) gedrucktem Text zu unterscheiden, noch Rotdruck durch Annotationen im Transkript abzubilden, wie dies etwa für Kursivdruck möglich wäre.

Da während der Laufzeit des Reichenau-Projektes keine öffentlichen Transkribus-Modelle für lateinische Inkunabeln vorlagen, liegt der Fortschritt durch die im Rahmen des Projektes trainierten Texterkennungsmodelle auf der Hand. Den nächsten Vergleichspunkt bildete das »Noscemus GM 6« Modell (Universität Innsbruck), das für die Erkennung wissenschaftlicher Drucke des 15.–19. Jahrhunderts erstellt wurde.[38] Dieses erzielt zwar für Antiquatypen sehr gute Ergebnisse, die Verwendung gebrochener Schriften für lateinische Texte liegt jedoch klar außerhalb seines Trainingshorizonts (vgl. Abb. 7). Ein quantifizierender Vergleich ist anhand der unterschiedlichen Transkriptionsansätze kaum möglich.

Umgang mit Forschungsdaten

Die automatisch erstellten Transkripte wurden bereits während der Projektlaufzeit in die Digitalen Sammlungen der Badischen Landesbibliothek eingepflegt und sind dort Seite an Seite mit den Digitalisaten einsehbar (vgl. Abb. 8).[39] Neben der Möglichkeit, den Text einzelner Sei-



8 Parallele Ansicht von Digitalisat und Volltext in den Digitalen Sammlungen der Badischen Landesbibliothek (Dares Phrygius, De excidio Troiae historia, Köln: Johann Schilling, bis 1472, BLB. Dg 46, fol. a1r)

https://doi.org/10.3196/1864295024713338 - am 13.01.2026, 14:16:01. https://www.inlibra.com/de/agb - Open Access -

ten in andere Programme zu kopieren, besteht hier auch die Option, mittel-auflösende Digitalisate ganzer Werke als PDF mit hinterlegter Textebene herunterzuladen.

Das im Rahmen des Projektes trainierte Texterkennungsmodell für lateinische Inkunabeln wird nach Ende der Projektlaufzeit (März 2024) innerhalb der Transkribus-Umgebung zur öffentlichen Nutzung freigeschaltet werden. Sämtliche Ground Truth-Daten des Projektes werden zusammen mit einer Dokumentation der zugrunde liegenden Transkriptionsrichtlinien unter einer CC-BY-SA-Lizenz in einem Forschungsdatenrepositorium dauerhaft zur Nachnutzung bereitgestellt und in der Datenbank der HTR-United Initiative verzeichnet.[40] Auf dieser Grundlage wird es möglich sein, äquivalente Texterkennungsmodelle für Inkunabeln auch außerhalb der Transkribusumgebung zu trainieren.

Fazit

Das in diesem Artikel vorgestellte Pilotprojekt demonstriert, dass eine qualitativ hochwertige Texterkennung historischer Drucke im Maßstab einiger zehntausend Seiten auch mit begrenztem Ressourceneinsatz realisierbar ist. Training und Anwendung eigener Texterkennungsmodelle benötigen keine spezialisierten IT-Kenntnisse; die hierbei erreichbare Qualität der Texterkennung ist auch für Inkunabeln sehr gut.

Die Unzulänglichkeiten automatischer Layouterkennung führen dazu, dass ein vollständig automatisierter Texterkennungsprozess als Voraussetzung einer wirklich »massenhaften« Volltexterschließung beim aktuellen Stand der Technik nur mit erheblichen Qualitätseinbußen umsetzbar wäre.[41] Die hierbei auftretenden Schwierigkeiten sind jeweils materialspezifisch und nicht vollständig antizipierbar. Aus den Erfahrungen dieses Pilotprojektes heraus empfiehlt es sich, vergleichbare Vorhaben eher im Sinne einer »computergestützten« als einer »automatischen« Volltexterschließung zu konzipieren und bereits zu Projektbeginn allgemeine Kriterien für die Abwägungen zwischen Automatisierungsgrad und Qualität bzw. Erschließungstiefe zu definieren.

Ausblick

Innovation im Bereich Texterkennung

Hinsichtlich möglicher Verbesserungen der Layoutanalyse auf historischen Drucken ist abzuwarten, welche Impulse sich aus dem Projekt »Robuste und performante Verfahren für die Layoutanalyse in OCR-D« (2023–2025, Staatsbibliothek zu Berlin / SLUB Dresden / ZPD Universität Würzburg)[42] ergeben werden.

Für die Texterkennung selbst bilden Transformerbasierte Modelle eine attraktive Option: Phillip Ströbel evaluierte die Anwendbarkeit TrOCR-basierter[43] Texterkennungsmodelle auf historische Handschriften und kam zu dem Schluss, dass diese eine weniger präzise Zeilensegmentierung als CNN×RNN Modelle erfordern, vor allem für uneinheitliches Material (z. B. mehrsprachige Texte) gut geeignet sind und auch mit geringem Trainingsaufwand eine gute Erkennungsqualität erreichen.[44] Transkribus integriert bereits einzelne Transformerbasierte Texterkennungsmodelle, erlaubt aber bislang nicht, diese materialspezifisch anzupassen (sog. Fine Tuning).

Die computergestützte Texterkennung handschriftlicher Marginalien, die historische Drucke zu Unikaten machen und so deren Mehrfachdigitalisierung rechtfertigen, rückt erst in jüngster Zeit in den Bereich des Möglichen:[45] Die geringe Menge an verfügbarem Textmaterial und die oft radikal abkürzenden Gebrauchsschriften der Schreiberinnen und Schreiber stellen Texterkennungssysteme vor einzigartige Herausforderungen, die aber mittelfristig lösbar erscheinen.

In jedem Fall legt der technische Fortschritt im Bereich der Texterkennung – 2015 etwa waren Zeichenfehlerraten von 5–8 % für Inkunabeln noch bemerkenswert[46] – nahe, die Volltexterschließung nicht als einmaligen Vorgang, sondern prozesshaft zu verstehen:[47] Werden gemessene oder geschätzte[48] Indikatoren zur Qualität der Volltexterkennung bereits im Rahmen des Erschließungsprozesses dokumentiert, lässt sich künftig besser entscheiden, welche Dokumente für die Anwendung verbesserter Texterkennungsmodelle, neuer Technologien oder auch als Objekte kollaborativer Korrekturinitiativen infrage kommen.

Anreicherung der Volltexte

Das Vorliegen qualitativ hochwertiger Volltexte eröffnet zahlreiche Möglichkeiten, die Digitalisate historischer Drucke besser lesbar zu machen und um weitere Informationen anzureichern: Zu denken ist etwa an das automatische Auflösen von Abbreviaturen, die Möglichkeit, Personen-, Orts-, und Objektnamen automatisch zu erkennen (Named Entity Recognition) und mit Normdatenbanken verknüpfen zu lassen (Named Entity Linking),[49] oder auch die automatisierte Übersetzung fremdsprachiger Texte.[50]

Bei der Planung solcher Vorhaben für Inkunabeln sind allerdings materialspezifische Schwierigkeiten zu berücksichtigen: Beispielsweise stellt die Auflösung von Wortumbrüchen am Zeilenende eine notwendige Vorbereitung für zahlreiche semantische Verarbeitungsverfahren dar. Während diese in modernen Drucken auf der Grundlage von Typografie (Trennstrichen) und Wörterbüchern problemlos möglich ist, stellen im Bereich der Inkunabeln eine unsystematische Verwendung von Trennzeichen, die oft den Erfordernissen des Blocksatzes geschuldet ist,[51] und die wenig standardisierte Orthografie erst noch zu bewältigende Herausforderungen dar.

Anmerkungen

1 Zur Sammlungsgeschichte vgl. HEINZER, Felix, 1989. *Die Reichenauer Inkunabeln der Badischen Landesbibliothek in*

https://doi.org/10.3196/1864295024713338 - am 13.01.2026, 14:16:01. https://www.inlibra.com/de/agb - Open Access -

Karlsruhe: Ein unbekanntes Kapitel Reichenauer Bibliotheksgeschichte. Sonderdruck aus: Bibliothek und Wissenschaft. 22, 1988. Wiesbaden: Harrassowitz. S. 24–80.

2 Vgl. Heinzer, 1989.

3 Wie vollständig dieses Korpus den historischen Inkunabelbestand der Reichenauer Klosterbibliothek wiedergibt, ist unsicher: Angaben des 19. Jahrhunderts gehen von über 200 Bänden aus. Vgl. Heinzer 1989, S. 15–19.

4 Vgl. Heinzer, 1989, S. 21–22.

5 Digitalisierung: 1.1.2023–31.12.2024, Volltexterschließung 1.4.2023–31.3.2024. Die zeitliche Verschiebung der beiden Projektkomponenten ergibt sich aus der Logik des Arbeitsablaufes (s. Abschnitt »Arbeitsablauf«).

6 Hauptziel des OCR-D-Projektes ist »die konzeptionelle und technische Vorbereitung der Volltexttransformation der VD [16, 17, 18]« (https://ocr-d.de/de/about).

7 Vgl. SPRINGMANN, Uwe, 2016. OCR für alte Drucke. In: *Informatik Spektrum*. 39(6), S. 459–462, hier S. 459 f., https://doi.org/10.1007/s00287-016-1004-3

8 Etwa Tesseract (https://github.com/tesseract-ocr) ab Version 4, kraken (https://kraken.re), Calamari (https://github.com/Calamari-OCR/calamari) und PyLaia (https://github.com/jpuigcerver/PyLaia).

9 https://gitlab.inria.fr/scripta/escriptorium, https://www.ocr4all.org/, https://www.transkribus.org/

10 Vgl. CHAGUÉ, Alix und Thibault CLÉRICE, 2023. Deploying eScriptorium online: notes on CREMMA's server specifications [online], 22.12.2023, https://alix-tz.github.io/phd/posts/017/

11 Vgl. https://readcoop.eu/coming-soon-new-transkribus-web-app/

12 Zu dieser Problematik vgl. PINCHE, Ariane und andere, 2023. CATMuS-Medieval: Consistent Approaches to Transcribing ManuScripts [online], 15.12.2023, hier S. 3, https://ehess.hal.science/hal-04346939v1

13 Vgl. insbesondere die Übersicht »Schreibweisen, spezielle Zeichen und Sonderzeichen« (https://ocr-d.de/de/gt-guidelines/trans/trSchreibweisen.html), sowie die entsprechenden Themenseiten im Wiki (https://github.com/OCR-D/ocrd-website/wiki/).

14 Zu CATMuS vgl. Pinche et al., 2023. Die für die CATMuS-Modelle verwendeten Transkriptionsrichtlinien sind für Handschriften in PINCHE, Ariane, 2022. Guide de transcription pour les manuscrits du Xe au XVe siècle [online], 16.6.2022, https://hal.science/hal-03697382 sowie für frühe Drucke in SOLFRINI, Sonia, Simon GABAY, Geneviève GROSS, Pierre-Olivier BEAULNES, Aurélia M. OLIVEIRA und Daniela SOLFAROLI CAMILLOCCI, 2023. Guide de transcription pour les imprimés français du XVIe siècle en caractères gothiques: Version A [online], 13.11.2023, https://hal.science/hal-04281804/ dokumentiert.

15 Für das Reichenauer Material erwiesen sich bestehende Richtlinien als lückenhaft bzw. nicht praktikabel: Beispielsweise bietet die OCR-D-Dokumentation keine Handreichung zur Abbildung des in lateinischen Texten allgegenwärtigen r-Hakens als Abkürzung für »ur«, wohingegen Pinche, 2022 ein Unicode-Zeichen empfiehlt (Combining Ur Above, U+1DD1), das in Standardschriftarten nicht darstellbar ist.

16 Wie etwa in den OCR-D-Richtlinien vorgesehen, vgl. https://ocr-d.de/en/gt-guidelines/trans/layout.html

17 Vgl. BASTIANELLO, Elisa, 2022. Digital Editions at the Bibliotheca Hertziana. In: *Journal of Art Historiography 27s* [online], Dezember 2022, hier S. 3, https://doi.org/10.48352/uobxjah.00004200 und RABUS, Achim & Aleksej TIKHONOV, 2022. How ›smart‹ is Transkribus in fact? Evaluating models with enhanced functionality [Aufzeichnung eines Vortrags bei der Transkribus User Conference 2022, Innsbruck, 30.9.2022]. 10.10.2022, https://www.youtube.com/watch?v=DdTj73MyeGg

18 Vgl. Rabus & Tikhonov, 2022, 13:36 min.

19 Vgl. z. B. den in CAMPS, Jean-Baptiste, Chahan VIDAL-GORÈNE und Marguerite VERNET, 2021. Handling Heavily Abbreviated Manuscripts: HTR engines vs text normalisation approaches, 2021. In: *Document Analysis and Recognition – ICDAR 2021 Workshops, Lausanne*, 2021. Cham: Springer, S. 306–316, https://doi.org/10.1007/978-3-030-86159-9_21 vorgestellten mehrstufigen Ansatz.

20 Vgl. GUÉVILLE, Estelle und David Joseph WRISLEY, 2023. Transcribing Medieval Manuscripts for Machine Learning: Version 3 [online], 3.10.2023, https://doi.org/10.48550/arXiv.2207.07726, S. 10–14.

21 Vgl. Guéville & Wrisley, 2023, S. 11, 13.

22 Die Character Error Rate (CER) ist als Summe der Einfügungen, Löschungen und Substitutionen definiert, die notwendig sind, um vom Transkript zum korrekten Text (Ground Truth) zu gelangen, geteilt durch die Anzahl der Zeichen in der Ground Truth. Vgl. STRÖBEL, Phillip, 2023. *Flexible Techniques for Automatic Text Recognition of Historical Documents* [Dissertation]. Zürich: Universität. 2023, https://doi.org/10.5167/uzh-234886, S. 64–66.

23 Von der Möglichkeit, im Zuge der Texterkennung eine Wortsegmentierung durchführen zu lassen, wurde im Rahmen des Projektes kein Gebrauch gemacht, da es sich als unmöglich erwies, die Textinformationen auf Zeilen- und auf Wortebene während der Bearbeitung in Transkribus synchron zu halten.

24 Zu Entwicklung und Elementen dieser typografischen Form vgl. DUNTZE, Oliver, 2005. Text und Kommentar in juristischen Drucken der Frühen Neuzeit. In: *Archiv für Geschichte des Buchwesens*. 59, S. 11–33.

25 Der BoW F1-Score kam zur Anwendung, da die CER für Unterschiede in Lesereihenfolge und Zeilensegmentierung anfällig ist, vgl. CLAUSNER, Christian, Stefan PLETSCHACHER und Apostolos ANTONACOPOULOS, 2020. Flexible character accuracy measure for reading-order-independent evaluation. In: *Pattern Recognition Letters*. 131, S. 390–397, https://doi.org/10.1016/j.patrec.2020.02.003, S. 390–391. Zur Berechnung vgl. STRÖBEL, 2023, S. 63–64; verwendet wurde die PRImA TextEval 1.5 Software (https://www.primaresearch.org/tools/PerformanceEvaluation).

26 10% der verfügbaren Ground Truth werden als Validierungsdatensatz (Validation set) nicht direkt ins Modelltraining miteinbezogen, sondern zu dessen Steuerung und Kontrolle verwendet. Da Transkribus die automatische Erstellung eines völlig separat gehaltenen Testdatensatzes nicht unterstützt, wird die Texterkennungsleistung auf den Validierungsdaten gemeinhin als Qualitätsmaß verwendet.

27 Vgl. NEUDECKER, Clemens, Karolina ZACZYNSKA, Konstantin BAIERER, Georg REHM, Mike GERBER und Julián Moreno SCHNEIDER, 2021. Methoden und Metriken zur Messung von OCR-Qualität für die Kuratierung von Daten und Metadaten. In: Michael FRANKE-MEIER, Anna KASPRZIK, Andreas LEDL und Hans SCHÜRMANN, Hrsg. *Qualität in der Inhaltserschließung*. Berlin: De Gruyter Saur. Bibliotheks- und Informationspraxis. 70, S. 137–165, hier: S. 145–146, https://doi.org/10.1515/9783110691597-009

28 Vgl. JANSSEN, Frans A., 2005. The Rise of the Typographical Paragraph. In: Karl A. E. ENENKEL und Wolfgang NEUBER, Hrsg. *Cognition and the Book: Typologies of Formal Organisation of Knowledge in the Printed Book of the Early Modern Period*. Leiden: Brill, Intersections. 4-2004, S. 9–32.

29 L̄c̄ 20–21 (Köln: Quentell, 1480; Köln: Quentell, 1481), L̄c̄ 22 (deutsche Ausgabe, Basel: Richel, 1481), L̄c̄ 25 (Straßburg: Prüß, nicht vor 1490).

https://doi.org/10.3196/1864295024713338 - am 13.01.2026, 14:16:01. https://www.inlibra.com/de/agb - Open Access -

30 Vgl. CHAMPION, Matthew S., 2017. *The Fullness of Time. Temporalities of the Fifteenth-Century Low Countries*. Chicago; The University of Chicago Press, S. 178–184.

31 Vgl. z. B. KIRCHNER, Felix, Marco DITTRICH, Phillip BECKENBAUER und Maximilian NÖTH, 2016. OCR bei Inkunabeln: Offizinspezifischer Ansatz der Universitätsbibliothek Würzburg. In: *ABI Technik*. 36(3), S. 178–188 https://doi.org/10.1515/abitech-2016-0036

32 Vgl. z. B. REUL, Christian, Christoph WICK, Maximilian NOETH, Andreas BUETTNER, Maximilian WEHNER und Uwe SPRINGMANN, 2021. Mixed Model OCR Training on Historical Latin Script for Out-of-the-Box Recognition and Finetuning. In: *Proceedings of the 6th International Workshop on Historical Document Imaging and Processing*. Lausanne, September 2021. New York: ACM, S. 7–12, https://doi.org/10.1145/3476887.3476910

33 Für das deutschsprachige und das bilinguale Texterkennungsmodell wurde jeweils das Transkribus Print M1-Modell (vgl. https://readcoop.eu/model/transkribus-print-multi-language-dutch-german-english-finnish-french-swedish-etc/) als Base Model verwendet, das auf deutlich mehr Trainingsmaterial basierende lateinische Modell profitierte nicht von der Verwendung eines generischen Base Models.

34 https://github.com/psb1558/Junicode-font. Die Autorin dankt Dorothee Huff (Tübingen) für den wertvollen Hinweis auf diese Schriftart.

35 Die angegeben Werte sind jeweils für die Texterkennung ohne Anwendung eines statistischen Sprachmodells berechnet. Ob dessen Verwendung sinnvoll ist, muss jeweils am Einzelfall geprüft werden: Im Fall der Inkunabeln führte die damit einhergehende Tendenz zur orthografischen Normierung oft zu schlechteren Ergebnissen.

36 Vgl. WICK, Christoph, Christian REUL und Frank PUPPE, 2020. Calamari: A High-Performance Tensorflow-based Deep Learning Package for Optical Character Recognition. In: *Digital Humanities Quarterly* [online]. 14(2), http://www.digitalhumanities.org/dhq/vol/14/2/000451/000451.html. Andere Transkribusmodelle für frühe Drucke erreichen Zeichenfehlerraten von 0,6–2 % auf ihren jeweiligen Validierungsdaten, z. B. Noscemus GM 6 (https://readcoop.eu/model/print-latin-texts-15-th-19th-htr/): 0,8 % CER, Noscemus GM 5: 0,6 % CER, SpanishGothic_XV-XVI_extended_v1.2 (https://readcoop.eu/model/spanish-gothic-15th-16th-century/): 0,9 % CER, Dutch_Romantype_Pylaia (https://readcoop.eu/model/dutch-romantype-print-16th-19th-century/): 1,4 %, Dutch_Gothic_Print_Pylaia (https://readcoop.eu/model/dutch-gothic-print-16th-18th-century/): 2 % CER.

37 https://github.com/qurator-spk/dinglehopper

38 https://readcoop.eu/model/print-latin-texts-15-th-19th-htr/

39 https://digital.blb-karlsruhe.de/Inkunabeln/topic/view/178827

40 Vgl. ROMEIN, C. Annemieke und andere, 2023. Exploring Data Provenance in Handwritten Text Recognition Infrastructure: Sharing and Reusing Ground Truth Data, Referencing Models, and Acknowledging Contributions: Version 3 [online], 24.3.2023, https://doi.org/10.5281/zenodo.8116009, S. 5–12 zu Best Practices für die Veröffentlichung von Ground Truth-Daten.

41 Je nach Art des Vorhabens kann dieser Qualitätsverlust akzeptabel sein: Ein Volltexterschließungsprojekt historischer Drucke mit über 1,3 Millionen Seiten wurde bspw. an der Bibliotheca Hertziana durchgeführt, vgl. Bastianello, 2022 und https://transkribus.humanitiesconnect.pub/

42 https://gepris.dfg.de/gepris/projekt/517459941

43 https://github.com/microsoft/unilm/tree/master/trocr

44 Vgl. Ströbel, 2023, S. 127–128.

45 Vgl. CHENG, Liang, Jonas FRANKENMÖLLE, Adam AXELSSON und Ekta VATS, 2024. Uncovering the Handwritten Text in the Margins: End-to-end Handwritten Text Detection and Recognition: Version 2 [online]. 29.1.2024, https://doi.org/10.48550/arXiv.2303.05929

46 Vgl. Kirchner et al. 2016, S. 186, Tabelle 2.

47 Vgl. BAIERER, Konstantin und Philipp ZUMSTEIN, 2016. Verbesserung der OCR in digitalen Sammlungen von Bibliotheken. In: *027.7 Zeitschrift für Bibliothekskultur*. 4(2), S. 72–83, https://doi.org/10.12685/027.7-4-2-155

48 Zu Methoden, die Qualität der Texterkennung zu schätzen bzw. zu prognostizieren, vgl. STRÖBEL, 2023, S. 107–124 und CLAUSNER, Christian, Stefan PLETSCHACHER und Apostolos ANTONACOPOULOS, 2016. Quality Prediction System for Large-Scale Digitisation Workflows. In: *12th IAPR Workshop on Document Analysis Systems (DAS)*. Santorini, 11.–14.4.2016, Washington, DC: IEEE Computer Society, S. 138–143, https://doi.org/10.1109/DAS.2016.82

49 Vgl. z. B. MENZEL, Sina und andere, 2021. Named Entity Linking mit Wikidata und GND: Das Potenzial handkuratierter und strukturierter Datenquellen für die semantische Anreicherung von Volltexten. In: Michael FRANKE-MEIER, Anna KASPRZIK, Andreas LEDL und Hans SCHÜRMANN, Hrsg. *Qualität in der Inhaltserschließung*. Berlin: De Gruyter Saur. Bibliotheks- und Informationspraxis. 70, S. 229–257, https://doi.org/10.1515/9783110691597-012

50 Vgl. z. B. FISCHER, Lukas, Patricia SCHEURER, Raphael SCHWITTER und Martin VOLK, 2022. Machine Translation of 16th Century Letters from Latin to German. In: *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, Marseille, 25.6.2022, Paris: European Language Resources Association, S. 43–50, https://doi.org/10.5167/uzh-218848

51 Vgl. JANSSEN, Frans A., 2017. The Graphic Design of the First Book Printed by Johann Schöffer (1503). In: Christoph RESKE und Wolfgang SCHMITZ, Hrsg. *Materielle Aspekte in der Inkunabelforschung*. Wiesbaden: Harrassowitz. Wolfenbütteler Studien zur Geschichte des Buchwesens. 49, S. 43–57, hier S. 56.

Verfasserin
Katharina Ost, Wissenschaftliche Mitarbeiterin (Projekt Nachlass Joseph von Laßberg, Abteilung regionalia), Badische Landesbibliothek (BLB), Erbprinzenstaße 15, 76133 Karlsruhe,
Telefon +49 721 175 2278,
ost.katharina@blb-karlsruhe.de
Foto: Valentin Marquardt

https://doi.org/10.3196/1864295024713338 - am 13.01.2026, 14:16:01. https://www.inlibra.com/de/agb - Open Access -